DOCUMENT RESUME

ED 124 596 TM 005 356

AUTHOR Steinheiser, Frederick H., Jr.
TITLE A Bayesian Simulation for Determining Mastery Calssification Accuracy.
PUB DATE [Apr 76]
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Classification; Computer Programs; Criterion Referenced Tests; Cutting Scores; *Decision Making; Mathematical Models; *Probability; Simulation; *Student Testing
IDENTIFIERS *Mastery Tests; Test Length

ABSTRACT

A computer simulation of Bayes' Theorem was conducted in order to determine the probability that an examinee was a master conditional upon his test score. The inputs were: number of mastery states assumed, test length, prior expectation of masters in the examinee population, and conditional probability of a master getting a randomly selected test item correct, and of getting an item incorrect. Classification accuracy was shown to be a function of all of the above parameters for any specified level of mastery (in the criterion-referenced sense). Specific results showed that for some combinations of prior information and test length, no information from the test could force a reversal in the decision rule, or provide classification accuracy within acceptable error bounds...hence, test results would be irrelevant. The vulnerability of a Bayesian model to changes in the prior probabilities was also demonstrated. For example, a 10% change in conditional probability was sufficient to completely reverse a classification rule across all test lengths studied, when the prior probability was held constant. Less drastic shifts occured with changes in the prior probabilities. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

Presented at the 1976 American Educational Research Association's Annual Meeting (Division D, Measurement and Research Methodology), San Francisco, April, 1976

7.14

ED124596

A BAYESIAN SIMULATION FOR DETERMINING MASTERY CLASSIFICATION ACCURACY

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Frederick H. Steinheiser, Jr.
Army Research Institute for the Behavioral and Social Sciences
Arlington, Virginia

IM005 356

Objectives: The educational decision-maker often wants to know if an examinee has mastered a sequence of instruction at some pre-specified level of acceptability. If the test score is above the minimal passing standard, the examinee may be classified as having mastered the instruction; if his score is below the minimal standard, he would be termed a "nonmaster" of the instruction. Because of many sources of variability, misclassifications are likely to occur, as shown in the following figure.

		True Competency State: Master	True Competency State: Nonmaster
Classification Based on Test Score	Master	True Positive	False Positive
	Nonmaster	False Negative	True Negative

The objective of the educational decision-maker is to maximize the true classifications (True Positives and True Negatives) and to minimize the false classifications (False Positives and False Negatives). The datum of interest is the conditional probability that a particular examinee is in a particular state of mastery, given his test score. The objective of the present paper is to examine the effect of such variables as test length, number of hypothesized mastery states, and the quality of the examinee population, on the probability that an examinee is in a particular state of mastery given his test score. Specifically, the following two questions were addressed: (a) What is the probability of (in)correctly classifying an examinee on the basis of his test score, and (b) How long must a test be, and what score is required so that classification decisions might be made with some specified lower limit of misclassification?

Theoretical framework: The statistical model which was used for classifying students into various mastery groupings, given their test scores, is based upon Bayes' Theorem, where:

$p(Mi|T)$ is the conditional probability of a particular student being classified as belonging in the ith mastery state given his test score; N is the test length; S is the number of mastery states hypothesized by the decision maker; $p(Mi|tj)$ is the conditional probability of a person in the ith mastery state getting the jth test item correct; $p(Mi)$ is the prior probability of the representation of the ith mastery state in the examinee population (the % of examinees who are estimated to be in the ith mastery state).

$$p(Mi|T) = \frac{\prod_{j=1}^{N} p(Mi|t_j)}{\sum_{i=1}^{S} p(Mi)^{N-1} \prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{N-1}}$$

It is assumed that the mastery states are mutually exclusive, the test items are of equal difficulty, that the test is a test of unitary skills, and that there is independence among items.

Methods and Techniques: A computer simulation of the Bayesian model was conducted using the following data:

(1) Test length (N) took on values of 5, 10, 20, 40 items;
(2) Number of hypothesized mastery states (S) varied from 2 to 3;
(3) Prior probability of mastery for a given examinee (P(M1)) took on values of .9, .7, .5 when two mastery states were assumed;
(4) Prior probabilities of mastery states 1, 2, and 3 took on values of .5, .3, and .2, respectively; and .25, .50, and .25, respectively, when three mastery states were hypothesized;
(5) Assuming two mastery states, the conditional probabilities of a master getting any single item correct took on the values of .9, .8, and .7; and for a nonmaster getting any single item correct, the values were .6, .5, and .4 (indicated by $p(1|Mi)$ in the Figures);
(6) Assuming three mastery states, the conditional probabilities of a master (M1), intermediate master (M2), and nonmaster (M3) getting any single item correct were .8, .6, and .5, respectively, and another set consisted of .9, .8, and .2, respectively ($p(1|Mi)$);
(7) The per cent correct observed scores took on the values of 60%, 70%, and 80%.

Data Source: The conditional probabilities in (5) and (6) were needed in order to obtain the values for the $p(Mi|tj)$ in the preceding formula. Along with an estimate of one of these conditional probabilities, it is assumed that the decision-maker could also supply an estimate of the prior probabilities for the states of mastery, the number of items on the test, and the number of mastery states. The only thing that he would observe is the per cent of the items that a given examinee got correct.

Results and Conclusions: Only a small portion of the results from the simulation can be described in the present abstract. Discussion must therefore be restricted to a case in which two states of mastery were assumed and the prior expectation of finding a master was equal to .9. The curvature of each line in Figure 1 shows how the probability of claiming that an examinee is a master given his test score changes as a function of test length, per cent correct observed, and conditional probabilities of a master and nonmaster getting any single item correct. (Additional graphs would show the effect of varying the prior expectation of mastery on $p(M|T)$). In this example, the prior expectation of finding a master in the examinee population is 90%. The conditional probabilities in A, B, C, and D show the probabilities of a master (M1) and nonmaster (M2) getting a typical item correct. Test length is plotted on the abscissa and the probability of the examinee's being a master (M1) given his observed test score (based upon % correct of the total test length) is plotted on the ordinate.

The effect of the test length variable on classification accuracy is dramatic: if the $p(M1|T)$ had to be at least .5 for a person to be called a master, then scores of 70% correct on a 10 item test would lead to a "mastery" classification. But a 70% score on a 20 item test would lead to a "nonmastery" classification. (Fig. 1A)

The effect of varying probabilities of a master making a correct response, p(correct|M1), can be seen by comparing graphs A, B, C, and D. For any test length or observed test score, the probability of being in the mastery state is greater in B than in A. This shift is most obvious for the 70% correct curve. Note that p(M1|T) on A for an observed score of 70% (28 out of 40 correct) is approximately .04. However, the p(M1|T) in B for 70% of a 40 item test correct is .87. The main reason for this abrupt change is the lowered requirement for mastery, from .9 to .8. The probability that ".9 persons" score only 70% on long tests is quite low, whereas for ".8 persons" the probability of scoring 70% is rather high. Graphs 1C and 1D illustrate further changes in the classification probability due to only .1 step changes in the probabilities of masters and nonmasters making a correct response.

The same data from Figure 1A can be used to answer the second question presented earlier: How long must a test be, and what score is required for classification decisions to be made with some specified lower limit of misclassification? Inspection of the curves in Figure 2 reveals that test length markedly influences classification accuracy. For the 40 item test, the region where p(M1|T) is greater than .1 and less than .9 extends from 71% to 77%. This means that the probability of misclassifying an examinee will exceed .10 only when observed scores range from 71% to 77% correct. In contrast, the region of the five item test for which p(M1|T) is greater than .10 and less than .90 extends from about 26% to about 79%. Hence, there is a much larger region for which the probability of misclassification exceeds .10. This procedure therefore shows what scores must be obtained so that a nonmastery decision could be made with at least 90% confidence; which, in effect, force a reversal in the prior beliefs of the decision maker.

Educational and scientific importance of the study: The Bayesian approach has been taken by others in devising methods for classifying examinees on the basis of test length and examinee qualities. However, the present version is less theoretically cumbersome, and gives a straightforward description of how classification accuracy is sensitive to the above variables. A general finding demonstrated by, but not necessarily limited to a Bayesian model, is that setting percentage cutoff scores as a means for defining mastery must take into account the test length. Classification accuracy is not invariant with percent correct. A specific result peculiar only to a Bayesian model is that classification accuracy is also a function of the qualities of the examinee population, or at least the decision-maker's estimates of those qualities. The model also allows confidence limits to be set for a given test when the examinee population qualities have been specified; these confidence limits then constrain the region of acceptable scores. Thus, if a region of misclassification error can be tolerated by the decision maker for a given population, the model specifies what the test length must be and what range of scores must be obtained in order to stay within the desired acceptable region.

FIG. 1.

P(M1) = .9 P(M2) = .1

60% Correct ———— 70% Correct —·—·—· 80% Correct — — — — —

A	B	C	D
P(1|M1) = .9	P(1|M1) = .8	P(1|M1) = .8	P(1|M1) = .7
P(1|M2) = .6	P(1|M2) = .6	P(1|M2) = .5	P(1|M2) = .4

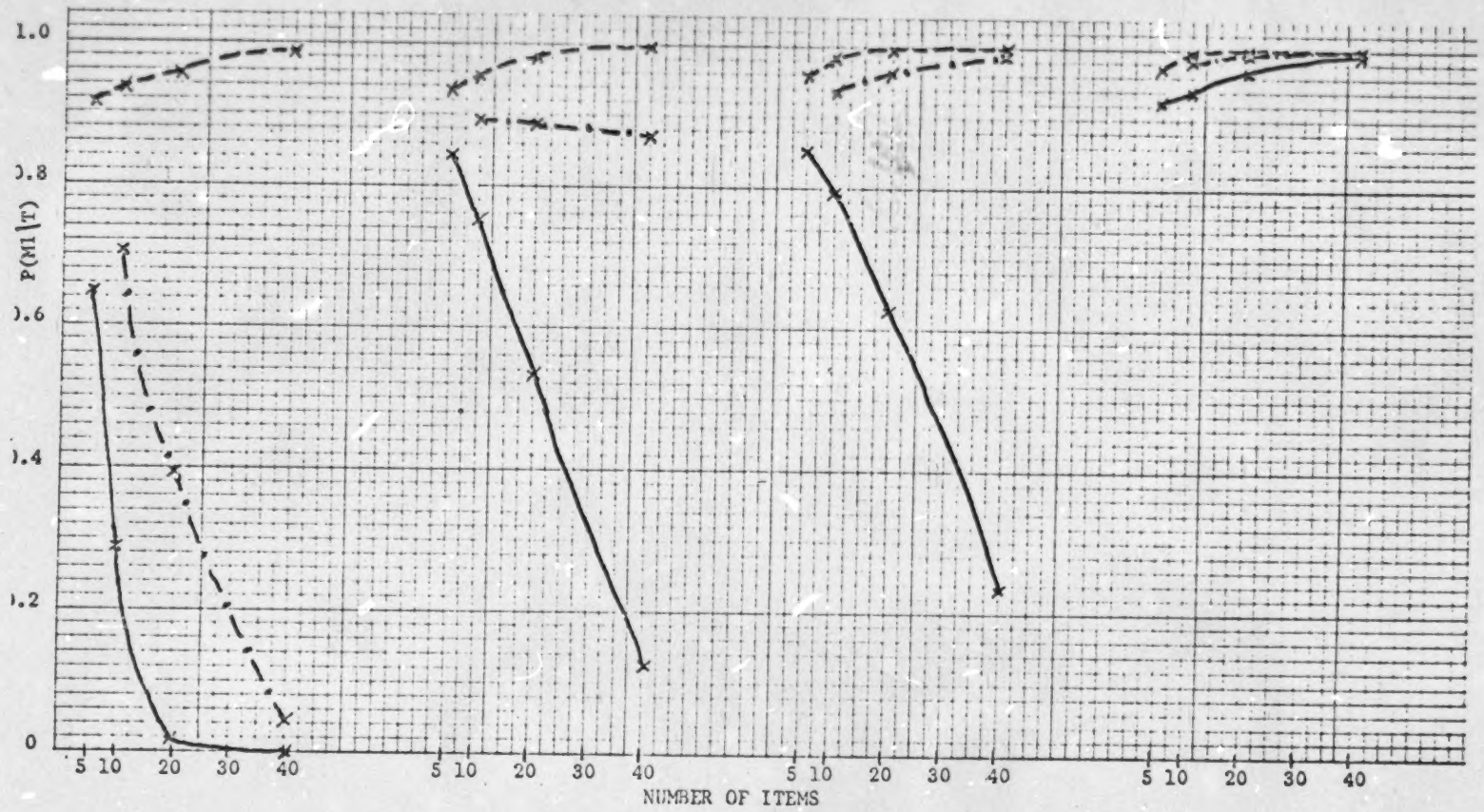


FIG. 2

P(M1) = .9 P(M2) = .1

P(1|M1) = .9, P(1|M2) = .6

5 Item Test – – – – – 40 Item Test ————

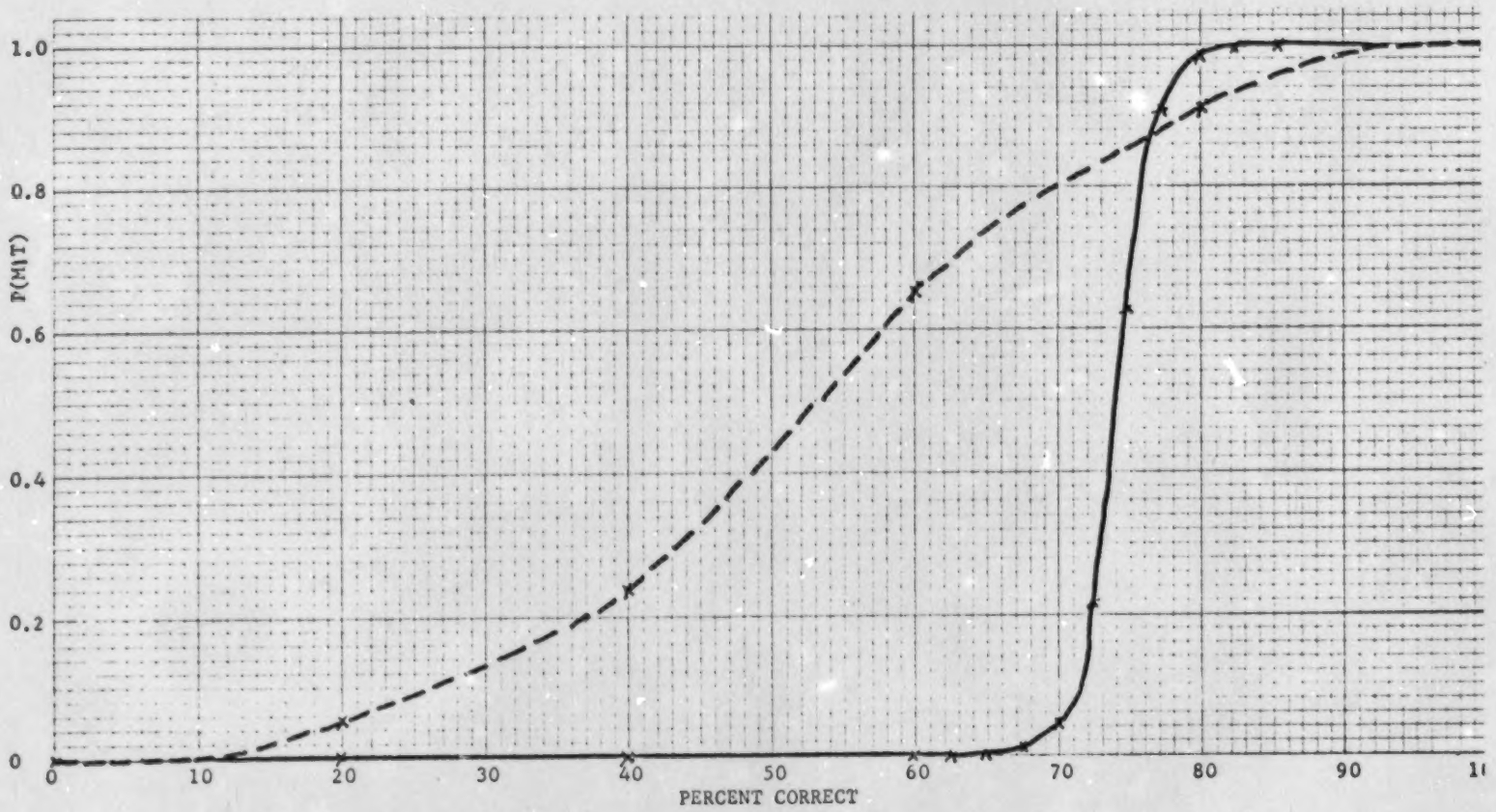


FIG. 3

P(M1) = .9 P(M2) = .1

P(1|M1) = .7, P(1|M2) = .4

5 Item Test — — — — — 40 Item Test ————

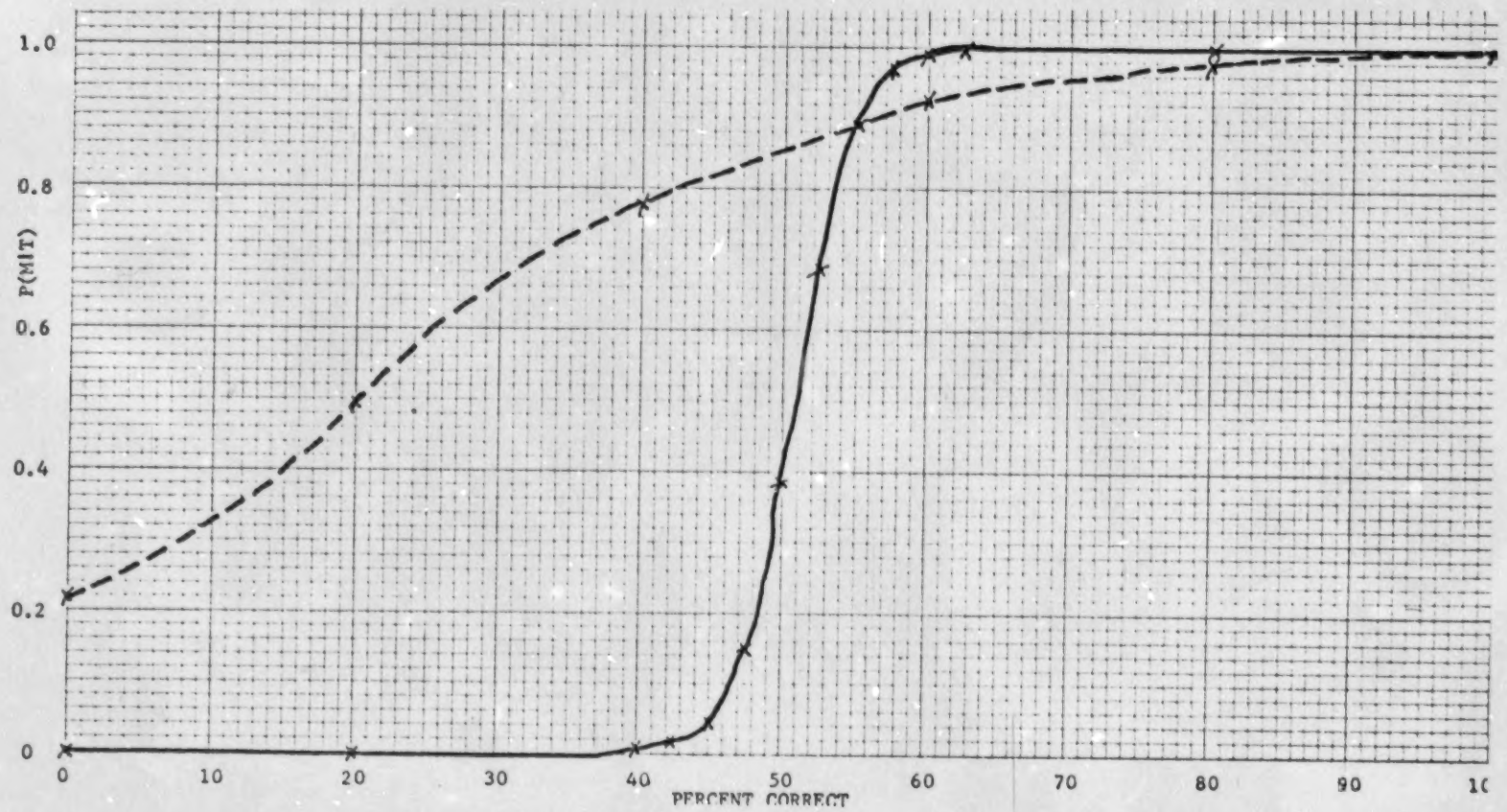
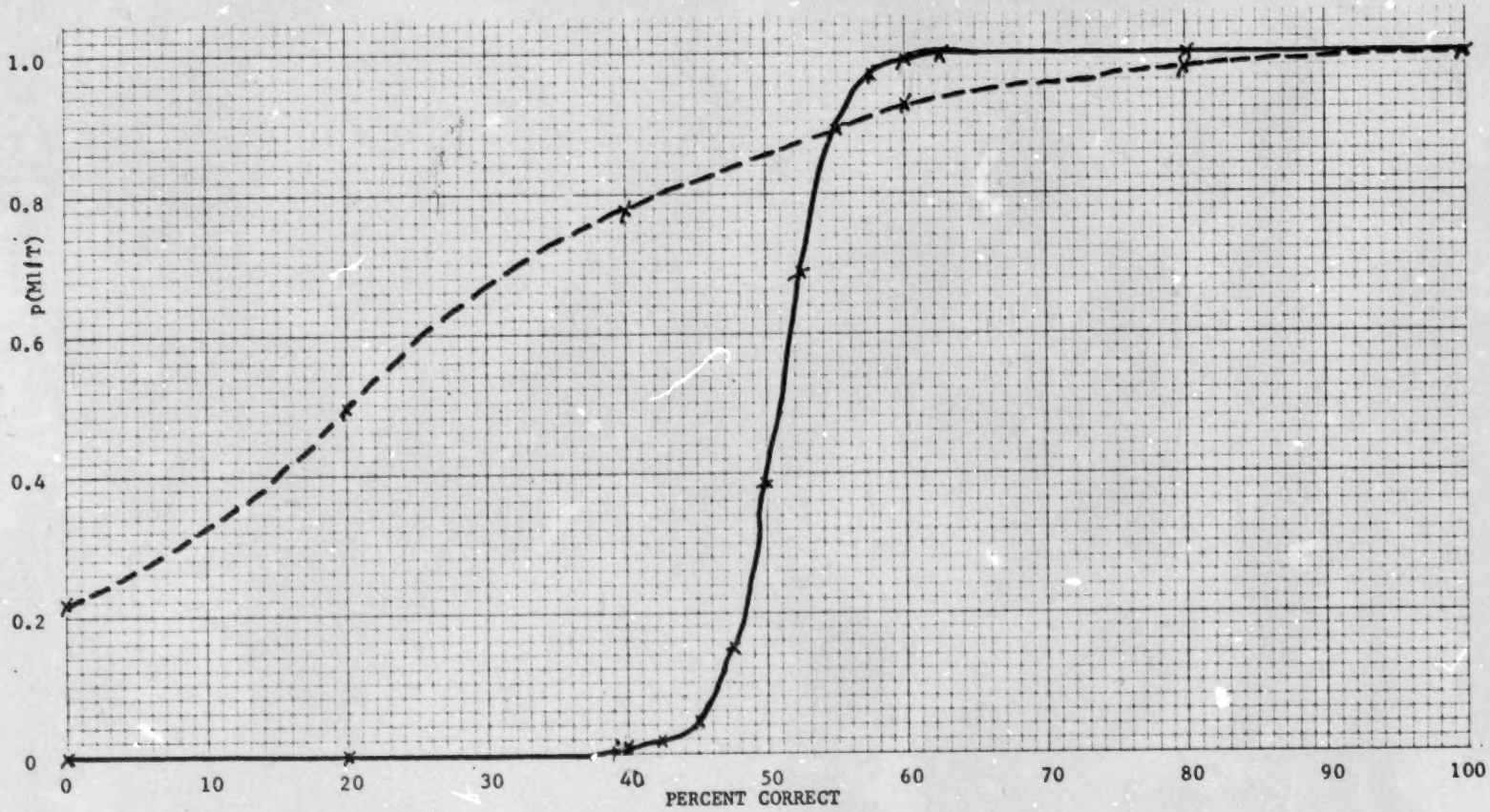


Fig 4

P(M1) = .9 P(M2) = .1

P(1|M1) = .7, P(1|M2) = .4

5 Item Test — — — — — 40 Item Test ———



Assume that there are three states of mastery, and unequal prior probabilities for these three states. The educational decision-maker must provide estimates for the prior probabilities of master, $p(Mi)$. For this example let us assume the values to be: $p(M1) = .5$; $p(M2) = .3$; and $p(M3) = .2$. He must also provide estimates for the conditional probability of getting any given test item right, given each mastery state. The following values will be used as the conditional probability of getting an item right given a mastery state: $p(1|M1) = .8$; $p(1|M2) = .6$; $p(1|M3) = .5$. The conditional probabilities of getting an item wrong given a mastery state are: $p(0|M1) = .2$; $p(0|M2) = .4$; and $p(0|M3) = .5$.

First we need to calculate the probability that an item is answered correctly. For the overall population, $p(tj = \text{correct})$

$$= \sum_{i=1}^{S} p(Mi)p(tj = \text{correct}|Mi) = (.5)(.8) + (.3)(.6) + (.2)(.5) = .68.$$

Likewise,

$$p(tj = \text{wrong}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{wrong}|Mi) = (.5)(.2) + (.3)(.4) + (.2)(.5) = .32.$$

We also need to obtain the set of conditional probabilities for the different mastery states given than an individual item was responded to either correctly or wrongly. The general equation is:

$$p(Mi|tj) = \frac{p(Mi)p(tj|Mi)}{p(tj)}.$$

Substituting the above values yields:
$p(M1|tj = \text{correct}) = (.5)(.8) \div .68 = .588$;
$p(M2|tj = \text{correct}) = (.3)(.6) \div .68 = .265$;
and $p(M3|tj = \text{correct}) = (.2)(.5) \div .68 = .147$.
(Note that the sum equals 1.0.) Finally,
$p(M1|tj = \text{wrong}) = (.5)(.2) \div .32 = .3125$
$p(M2|tj = \text{wrong}) = (.3)(.4) \div .32 = .375$ and
$p(M3|tj = \text{wrong}) = (.2)(.5) \div .32 = .3125$
If 6 items were answered correctly on a 10 item criterion-referenced test, the following $\prod_{j=1}^{N} p(Mi|tj)$ values result:
$M1 = 3.9 \times 10^{-4}$; $M2 = 6.8 \times 10^{-6}$;
$M3 = 9.6 \times 10^{-8}$
Finally, the general Bayesian formula yields the conditional probability for each mastery state given the total test score. For example, $p(Mi|T) =$

$$\frac{(3.9 \times 10^{-4})}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(6.8 \times 10^{-6})}{(.3)^9} + \frac{(9.6 \times 10^{-8})}{(.2)^9}\right]}$$

$= .272$.
Similar calculations yield $p(M2|T) = .473$ and $p(M3|T) = .254$.

In order to combine mastery states M2 and M3 into a single mastery state (which could represent combining the two degrees of nonmastery, Figure 4, Graph D), the following calculations are required. The values for $\prod_{j=1}^{N} p(M1|tj)$ remain the same, .5 and 3.9×10^{-4} respectively. The new nonmastery state (M2') occurs as a result of combining the previous states M2 and M3. Hence, $p(M2') = p(M2) + p(M3) = .3 + .2 = .5$, $p(M2'|tj = \text{correct}) = p(M2|tj = \text{correct}) + p(M3|tj = \text{correct}) = .265 + .147 = .412$, and $p(M2'|tj = \text{wrong}) = p(M2|tj = \text{wrong}) + p(M3|tj = \text{wrong}) = .375 + .3125 = .6875$.

Calculation of $\prod_{j=1}^{N} p(M2'|tj)$ yields
1.09×10^{-3}.

Entering these new values into the general Bayesian Formula, the following values of $p(M1'|T)$ and $p(M2'|T)$ are obtained:

$$p(M1'|T) = \frac{3.9 \times 10^{-4}}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9}\right]}$$

$= .264$,

$$p(M2'|T) = \frac{1.09 \times 10^{-3}}{(.5)^9\left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(1.09 \times 10^{-3})}{(.5)^9}\right]}$$

$= .736$.

Fig 5

P(M1) = .50 P(M2) = .30 P(M3) = .20

60% Correct ——————— 70% Correct —·—·—·— 80% Correct — — — —

P(1|M1) = .8, P(1|M2) = .6, P(1|M3) = .5